Europäisches Patentamt	European Patent Office	Office européen des brevets

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

00480041.3

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

I.L.C. HATTEN-HECKMAN

DEN HAAG, DEN
THE HAGUE, 03/08/00
LA HAYE, LE

Europäisches Patentamt
European Patent Office
Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.:
Application no.:
Demande n°: 00480041.3

Anmeldetag:
Date of filing:
Date de dépôt: 12/05/00

Anmelder:
Applicant(s):
Demandeur(s):
INTERNATIONAL BUSINESS MACHINES CORPORATION
Armonk, NY 10504
UNITED STATES OF AMERICA

Bezeichnung der Erfindung:
Title of the invention:
Titre de l'invention:
System and method of uniquely authenticating each replication of a group of soft-copy documents

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat: State: Pays:	Tag: Date: Date:	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation:
International Patent classification:
Classification internationale des brevets:
/

Am Anmeldetag benannte Vertragstaaten:
Contracting states designated at date of filing: AT/BE/CH/CY/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du depôt:

Bemerkungen:
Remarks:
Remarques:

SYSTEM AND METHOD OF UNIQUELY AUTHENTICATING EACH REPLICATION OF A GROUP OF SOFT-COPY DOCUMENTS

Field of the Invention

The present invention relates to the field of document authentication and more particularly applies to a system and method in which one soft-copy document, out of a group of associated documents, acts as the carrier of the integrity information aimed at uniquely authenticating each replication of the group.

Background of the Invention

The current environment of computer networks is characterized by an exponential growth in the circulation of soft-copy or electronic documents. They include plain text documents or text-like document e.g., ASCII (American Standard Code for Information Interchange) coded files and more generally data files such as the ones corresponding to the coding of images e.g., according to the JPEG (Joint Picture Expert Group) standard compression algorithm. However, because this has to take place over unsecured media especially, the Internet, a key issue becomes authentication. It should be possible for the recipient of a document to ascertain its origin so that no one should be able to masquerade as someone else. Also, it should be possible to verify that a document has not been modified, accidentally on maliciously, en route. To this end the standard solution; which goes well with any form of electronic document since, whatever method is used to code the information, the end result is just a binary data file, consists in concatenating a MAC or Message Authentication Code to the soft-copy document files. A MAC is a digest computed with a one-way hash function over a coded binary file, which is also made dependent on a key e.g., a secret-key known only to the sender and the receiver in order this latter can check first, that what it received has indeed been originated by whom shares the secret-key with it and second, that the document has not been altered. For example, Secure Hash Algorithm or SHA specified by the National Institute of Standards and Technologies, NIST, FIPS PUB 180-1, "Secure Hash Standard", US Dpt of Commerce, May 93, produces a 160-bit hash. It may be combined with a key e.g., through the use of a mechanism referred to as HMAC or Keyed-Hashing for Message Authentication, subject of the RFC (Request For Comment) of the IETF (Internet Engineering Task Force) under the number 2104. HMAC is devised so that it can be used with any iterative cryptographic hash function thus, including SHA. Therefore, a MAC can be appended to a document file so as the

whole can be checked by the recipient. Thus, this method assumes the addition of checking information to an existing file after the information to be transmitted has been coded. This has the inconvenience of indeed clearly separating the file content information from its checking part. Hence, this latter can easily be isolated and removed intentionally, in an attempt to cheat, or accidentally just because the intermediate pieces of equipment, in charge of forwarding the electronic documents, are not devised to manipulate this extra piece of information.

Yet another key issue with a public communications network such as the Internet, is privacy and confidentiality. Not all the information circulating between end users, be it comprised of texts, images or is a combination of, should be made public. The standard answer to this issue rests on cryptography. That is, information files that must be kept secret are encrypted before transmission. DES (Data Encryption Standard) is the standard encryption algorithm that has been in use for two decades to encrypt and decrypt data files. It operates on 64-bit blocks of data, using a symmetric secret-key to be shared by those involved. DES is identical to the ANSI standard Data Encryption Algorithm (DEA) defined in ANSI X3.92-1981.

Authenticating encrypted files is conducted just as with non-encrypted files i.e., a MAC is computed and concatenated to what remains intrinsically a binary data file. Hence, another disadvantage of computing integrity information on data is that the integrity information itself reveals some information about the data on which it is computed. Unless the key is changed, the integrity information computed on the data remains constant. Therefore, if an eavesdropper observes the same transmitted MAC he/she can be certain that the same encrypted message was transmitted. In applications where pre-formatted files are repeatedly forwarded (e.g., the same coded images or coded pieces of music), a simple frequency analysis performed on the intercepted MAC values may reveal a pattern in the

transmitted messages. Similarly, one or a set of encrypted files sent only once however, to many destinations, just unveils what group has received the same information. Hence, it would be advantageous firstly, to introduce randomization into the process so that MAC values are constantly changing and secondly, to allow the hiding of MACs in the transmitted information thereby, completely preventing an adversary from learning anything through the observation of the forwarded data. If this result could be partially obtained (randomization could be achieved this way) by changing the keys especially, the key to compute the MAC, that would have to be different for each transmitted copy of identical data file(s) and/or for each different destination, this would however severely impacts the key management system since this would assume that secret-keys be first distributed through a separate channel, a burdensome task.

Object of the Invention

Thus, it is a broad object of the invention to remedy the shortcomings of the prior art as described here above.

It is another object of the invention to disclose a method and a system which introduce randomization in the insertion of a MAC so as, to each replication of a same group of data files, unique authentication data can be associated.

It is still another object of the invention to allow that authentication data be merged and hidden in the transmitted information.

It is a further object of the invention of being transparently implement able especially, without the need of having to distribute more secret-keys than usually required to perform encryption and authentication.

Further objects, features and advantages of the present invention will become apparent to the ones skilled in the art upon examination of the following description in reference to the accompanying drawings. It is intended that any additional advantages be incorporated herein.

Summary of the Invention

A method and a system for uniquely authenticating each replication of a plurality of soft-copy documents, forming a group, are disclosed. The method first selects one soft-copy document, out of the group, to become a carrier for an authentication code aimed at protecting the group. The soft-copy documents are concatenated however, using a canonical form of the selected soft-copy document so as checking can be later successfully performed. Then, the authentication code is computed from the concatenation of the plurality of soft-copy documents and a key. A random number is also generated for each replication of the group of files. It is combined with the authentication code to mark the selected soft-copy document so as to obtain the carrier. A method for checking the authentication code is also disclosed.

Thus, the invention introduces randomization in the insertion of an authentication code so as, to each replication of an identical group of data files, unique authentication data can thus be associated. Invention also allows to merge and hide them in the transmitted information.

Brief Description of the Drawings

Figure 1 is a simple example of how the invention can be carried out.

Figure 2 shows that more files can be associated in a group.

Figure 3 describes the general case where a plurality of files, including the carrier, are associated.

Figure 4 is a detailed description of how the invention is carried out when the carrier is a plain text file.

Figure 5 shows the step of the checking method per the invention.

Figure 6 shows an example of an improved method to obtain an authentication code per the invention.

Detailed Description of the Preferred Embodiment

Figure 1 illustrates how the invention can be better carried out through an exemplary application. A first document, here a picture of a fingerprint [100], is encoded into a data file formatted e.g., according to the JPEG (Joint Photographic Experts Group) standard image compression algorithm thus, producing a data file [110] name of which is, for example, fingerprint.jpg. Associated to this image there is a second document, for example a text [120], giving some explanations on what are fingerprints. Text document is also coded into a file e.g., a simple ASCII file 'fingerprint.txt' [130]. Then, first document, the fingerprint image [100], can be protected by computing [140] a MAC (Message Authentication Code) using the file message [110] and a key [115] as inputs through any method well-known from the art thus, obtaining a unique digest or MAC [145] i.e., a binary vector made of 1's and 0's. Then, instead of appending the MAC to the fingerprint file it is rather used, along with a randomly generated number [160] to transparently mark [170] in a unique way, the image companion text file [130], as explained in following figures, in order to fulfill the objectives of the invention that are to randomize and hide the authentication data. Therefore, the second document acts as a carrier [150] for the authentication data.

Figure 2 just illustrates the fact that more than two documents can be involved. As an example, a picture of the person [205] whose fingerprint is shown [200] can also be associated so that a MAC is computed [240] over the concatenated picture files [210, 211] then merged into the text file [230] so as both can be together authenticated through the decoding of the marked companion text [250] upon reception of the three associated files [210, 211, 250].

Figure 3 further illustrates a general case for the invention assuming that MAC [345] is not only computed over more than one file e.g., the files [310, 311] of the two images [300, 305] but also includes the file of the text document itself [320], that is eventually used to carry [350] the authentication data, so all three pieces form a group [313] that can be authenticated together. Because, in this case, the carrier of the authentication data is concatenated [314] so as it participates into the computation [340] of the MAC it is implicitly assumed that it may exist such a thing as a canonical form of the carrier and text document file [330] from which an identical file [312] (canonical.txt) may be generated, in order that computation performed at generation and at checking can indeed match. In this particular example, since the carrier is the text-document file [330], the modifications mentioned here above that may be brought to the text document and are transparent should consist in changing, in one way or another, the number of inter-word blank characters (like e.g., [331]) of the text. This does not affect readability whatsoever. Then, a canonical form of the text that must be agreed upon by all parties involved, from which MAC computation must start, may consist in removing all inter-word blank characters or alternatively leaving a predefined fixed number of those blank characters say, one blank between any two words as it is usual in a text document. Thus, canonical form is, in this particular case, a form of text that can be obtained identically by sender and receiver irrespective of the fact that text has been marked (through the insertion of extra blanks) so as to permit authentication of the set of transmitted files.

Finally, like with other examples of figure 1 and 2 MAC is used, along with a randomly generated number [360] to transparently mark [370] the text file [130] which a carrier [350] for the authentication data.

Figure 4 describes the method of the invention to permit the embedding, into the carrier file, of the authentication data so as the information is randomized and hidden. Although this preferred embodiment of the invention is exemplified with the use of a text file, as a carrier for the authentication data, it should be clear to those skilled in the art that, without departing from the spirit of the invention, it could be practiced in many different ways that may, or may not, involve a text file. Carrying out the invention assumes it exists some sort of neutral element in the carrier file which does not alter its meaning or function if present alone or replicated in excess of what is strictly necessary. In accordance with this definition, blank character (x'40' for an ASCII file, here displayed with a caret sign ^ e.g., [480]) is the neutral element for a text document since inserting more blanks than necessary i.e., one blank, does not alter readability. Also, because there is at least one blank between any two words, there are many opportunities for merging, under the form of extra inter-word blanks, the authentication data if the text is indeed comprised of enough words.

Then, invention assumes that MAC [400] (computed according to any standard or custom method known from the art) is used here to split the text into two sets of words. This is simply achieved by creating a first set [410] with the words whose position correspond to the 1's [412] of the MAC binary vector. The second set [420], the complement, corresponds to the 0's [422]. It is assumed in the description of figure 4, for the sake of simplicity, that the length of the MAC binary vector matches the number of words of the text [430] even though that may seldom be the case. However, if the text is comprised of more words, the common case, all parties involved should agree on what part is to be selected to carry out encoding and checking similarly. Although the simplest method would be to consider that enough leading words, to match the length of the MAC binary vector, should be selected many other alternatives are possible like selecting the trailing words

instead or any other more sophisticated way of selecting the subset of words which must be agreed on beforehand. It is also assumed that the number of words of the shortest carrier text must be large enough to match the length of the binary vector result of the chosen MAC function. In other words, depending on the level of protection one desires to achieve in a particular application of the invention, the binary MAC vector will have different lengths (e.g., a 160-bit hash is produced with SHA). Thus, carrier text must have at least the corresponding number of words (more exactly, there must be enough inter-word intervals thus, generally excluding the last word of a text [431]) to permit the use of the complete MAC binary vector. Otherwise, if MAC cannot be entirely used, because carrier text is too short, the level of protection is reduced accordingly.

Then, the first set of words [410] along with their trailing inter-word blank characters, is marked through the insertion of a random number (RN) of extra blank characters in it. The exact method to achieve this is beyond the scope of the invention. For the sake of clarity, a straightforward way is assumed in this illustration of the invention. That is, RN is generated from any convenient random number generator [445] known or adapted from the art, under the form of a first binary vector P1 [440] e.g., fitting into the smallest of the two sets of words. Then, one extra blank is inserted for each interval corresponding to a 1 in P1 e.g., [442]. Although more sophisticated methods and numerous variations for obtaining and inserting a random pattern of extra blanks can be considered this would not change the scope of the invention which rather than imbedding directly the MAC into the carrier text uses it to split it and insert a random number of extra blanks instead.

Finally, the second set of words is marked too. That is, starting from P1 [440], a transform function T [450] is applied to get a second pattern P2 = T(P1) [460] that must fit into this second set of words. Similarly to P1, P2 is used to

insert extra blanks into the second set e.g., [462]. Again, the type of transform T to be used to obtain P2 from P1 is beyond the scope of the invention. Many equivalent alternates ways are possible. The simplest one consists in just reusing P1 as is to fit in the second set. A more elaborated transform is to hash RN so that P2 = H(P1). H would be any appropriate hashing function. Here P2 [460] is just the bit-wise complement of P1. The last operation consists in reassembling [470] the two sets from the pattern of 1's and 0's of the MAC [400] thus, obtaining the carrier [475].

Figure 5 describes an example of the decoding method per the invention when carrier is a plain text document in which extra blanks have been inserted. Description starts when a MAC is re-computed [500] as explained in figure 1, 2 and 3. The operations shown in this figure are effected in a manner similar to what was shown in figure 4. Then, with the MAC binary vector, text is split in a first and a second set. From both sets a pattern of extra blanks is extracted [520, 530]. On the pattern extracted from the first set the same transform, as used for encoding, is applied [540]. Finally, if the pattern of extra blanks of the second set matches [551] the transformed pattern of the first set when compared [550] then files are accepted as authentic. If comparison fails [552] one or more files of the group of files should be considered as having been compromised.

Figure 6 elaborates on what should be the structure of the MAC to best carry out the invention. Although the invention does not require to make any assumption on the manner a MAC is calculated it is however worth mentioning the following restrictions. Whichever method is actually retained for computing it, the case where a MAC having very few 1 bits or very few 0 bits must however be considered, even though the probability of obtaining such a ratio of 0's and 1's is low or very low. Hence, let's assume for a moment that, e.g., a

128-bit MAC has 127 one bits and only one 0 bit. In effect, that would result in 127 bits of random data that could be encoded into the first set of words (such as [410] in figure 4) of the split text and only 1 bit of transformed random data could be encoded into the second set (such as [420]) thus, actually comprised of only one word of the text in this extreme example. In that case, an adversary who substitutes a different text for the real text, would have a probability of 1/2 of accidentally passing the verification check. Similarly, if we have a split of 126 and 2 there would be a probability of 1/4 of accidentally passing the verification check and so on. Therefore, the optimal situation is when the MAC has the same number of 1's and 0's i.e., 64 ones and 64 zeros in this example of a 128-bit MAC. That yields the minimum probability of 1/(2**64). As a consequence, one may optionally want to favor a method, for generating MACs, that warrants a prescribed number of 0 and 1 bits. Among numerous possibilities, one trivial method is to have a recursive procedure where the MAC is computed on the input data [600], a key [610], and a counter [620]. The generated MAC [630] is tested [640] to see whether the number of 0 bits and 1 bits satisfies the imposed condition, and if so then the MAC is accepted and used [650]. Otherwise, the counter [620] is incremented [660] and a new MAC is computed, and this procedure continues until an acceptable MAC is found. Because sender and receiver use the same procedure to generate MACs and since the first acceptable MAC is taken, they are indeed assured to make use of the same MAC value.

Claims:

What is claimed is:

1. A method for uniquely authenticating each replication of a plurality of soft-copy documents [310, 311, 330], said plurality of soft-copy documents forming a group [313], said method comprising the steps of:;

selecting one soft-copy document [330], out of said group [313], to become a carrier [350] for an authentication code [345] aimed at protecting said group;

concatenating [314] said plurality of soft-copy documents said step of concatenating including the step of:

using a canonical form [312] of said selected soft-copy document [330];

computing [340] said authentication code [345] from said concatenated plurality of soft-copy documents [314] and a key [315];

generating a random number [360];

combining said random number and said authentication code to mark [370] said selected soft-copy document [330];

thereby, obtaining said carrier [350].

2. The method according to claim 1 wherein said step of concatenating omits said selected soft-copy document [230].

3. The method according to any one of the previous claims wherein said step of concatenating is replaced by the step of picking up a single soft-copy document [110].

4. The method according to any one of the previous claims wherein the combining step includes the steps of:

splitting said selected soft-copy document [430] into a first set [410] and a second set [420] on the basis of said authentication code [400];

utilizing said random number [440] to mark [442] said first set;

transforming [450] said random number;

utilizing said transformed random number [460] to mark said second set;

reassembling [470] said first set [410] and said second set [420] on the basis of said authentication code [400];

thereby, obtaining said carrier [475].

5. The method according to any one of the previous claims wherein said selected soft-copy document is a plain text document [430] and said first set [410] and said second set [420] are sets of words from said plain text document.

6. The method according to any one of the previous claims wherein said plain text document [430] is marked through the insertion of extra blanks [442] [462].

7. The method according to any one of the previous claims wherein the step of using said canonical form of said plain text document includes the step of:

stripping all interword blank characters, in excess of one, off said plain text document;

thereby, obtaining said canonical form.

8. The method according to any one of the previous claims wherein said authentication code [400], said random number [440], said transformed random number [460] are binary vectors fitting respectively in said selected soft-copy document [430], said first set [410] and said second set [420].

9. The method according to any one of the previous claims wherein said splitting step includes the steps of:

forming said first set with the words from said selected soft-copy document corresponding to the ones of said authentication code [412];

forming said second set with the words from said selected soft-copy document corresponding to the zeros of said authentication code [422];

10. The method according to any one of the previous claims wherein said computing step is replaced by the steps of:

computing [605] said authentication code [630] from said concatenated plurality of soft-copy documents [600], said key [610] and a counter [620];

testing [640] one's density of said authentication code;

if failing test:

incrementing [660] said counter;

resuming at computing step;

if passing test:

validating [650] said authentication code;

exiting said computing step.

11. The method according to any one of the previous claims wherein said transforming step includes:

hashing said random number;

reusing said random number;

inverting said random number.

12. A method for checking said authentication code comprising the steps of:

obtaining [500] said authentication code;

splitting [510] said carrier into a said first set and a said second set;

extracting [520] a first pattern from said first set;

transforming [540] said first pattern;

extracting [530] a second pattern from said second set;

comparing [550] said transformed first pattern and said second pattern;

if matching [551]:

passing checking;

if not matching [552]:

failing checking.

13. A system, in particular a system for uniquely authenticating each replication of a group of soft-copy documents, comprising means adapted for carrying out the method according to any one of the previous claims.

14. A computer-like readable medium comprising instructions for carrying out the method according to any one of the claims 1 to 12.

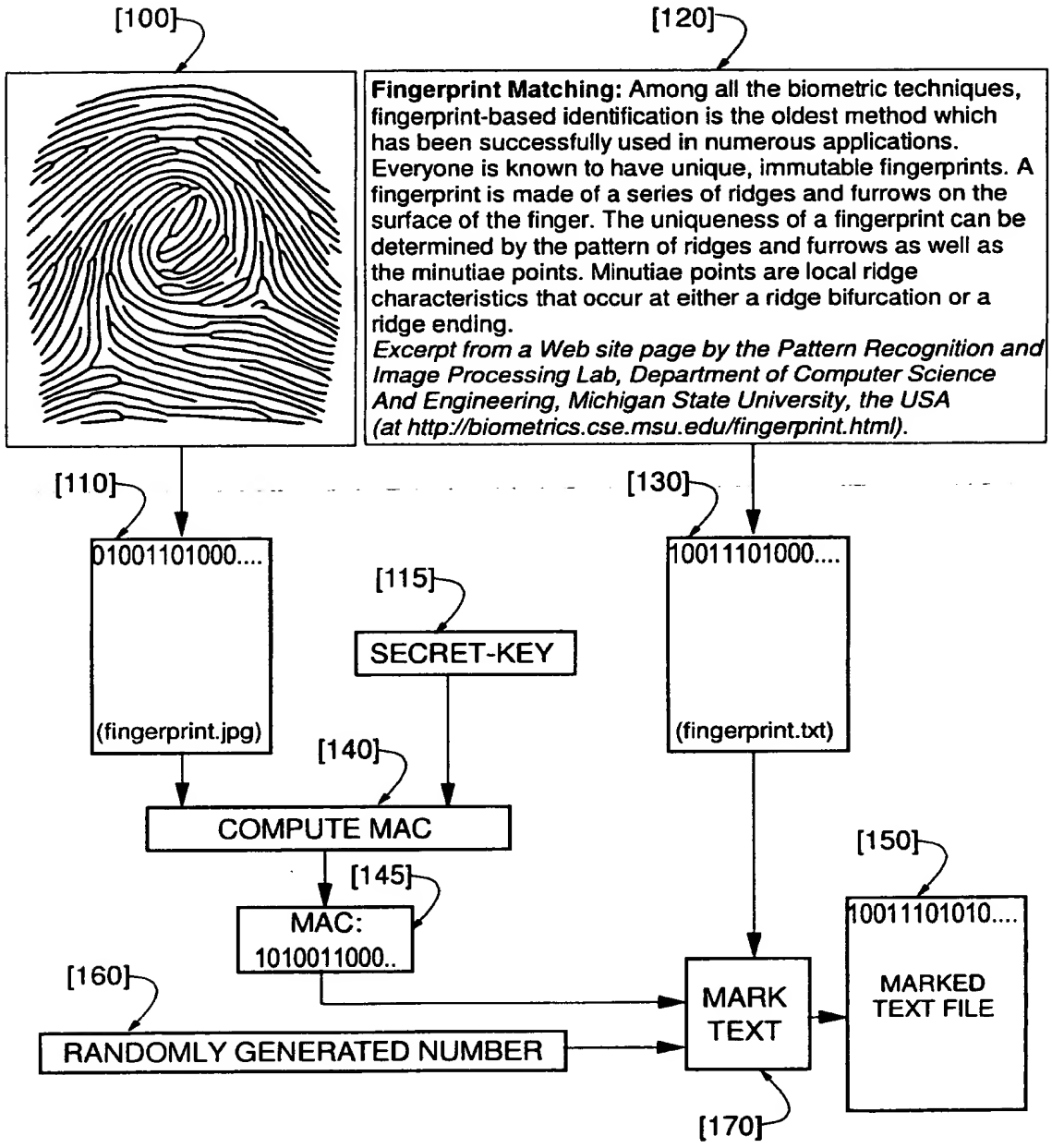
SYSTEM AND METHOD OF UNIQUELY AUTHENTICATING EACH REPLICATION OF A GROUP OF SOFT-COPY DOCUMENTS

Abstract

The invention allows an unique authentication of each replication of a plurality of soft-copy documents, forming a group. One soft-copy document, out of the group, becomes a carrier for an authentication code aimed at protecting the group. The authentication code is computed from the concatenation of the plurality of soft-copy documents and a key. A random number is also generated for each replication of the group of files. It is combined with the authentication code to mark the soft-copy document which has been selected to become the carrier. Thus, the invention introduces randomization in the insertion of the authentication code so as, to each replication of an identical group of data files, unique authentication data can be associated also, further merged and hidden in the transmitted information.

Figure 3.

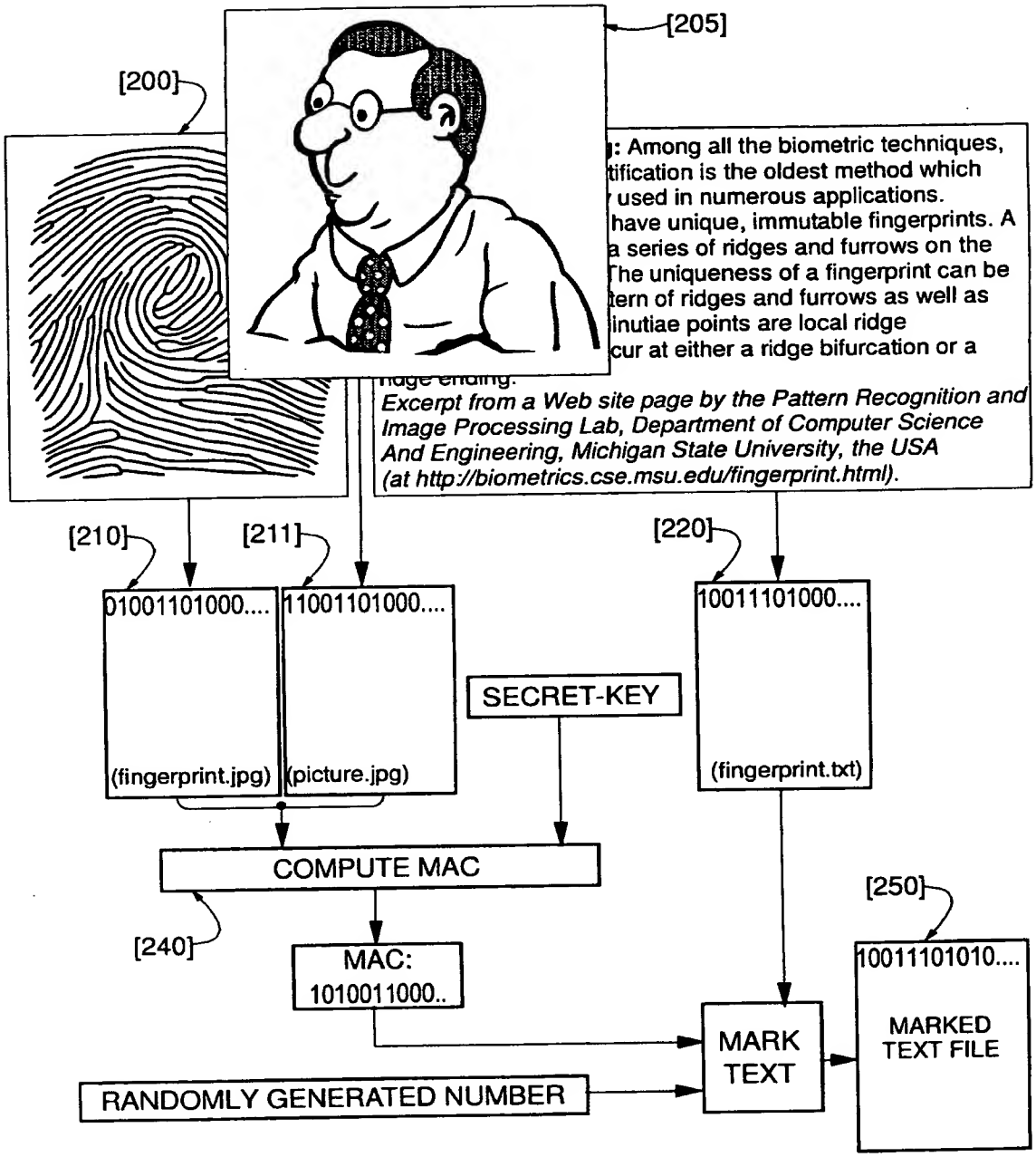
FR 9 2000 0021
INCERTIS CARRO et al.
1/6



Note: fingerprint image [100], used to illustrate the invention, is the vectorization of one image out of a 36-image animated GIF file borrowed from a Web site by the Biometric Systems Lab, University of Bologna, Cesena, Italy at http://www.csr.unibo.it/research/biolab/bio_tree.html

Figure 1

FR 9 2000 0021
INCERTIS CARRO et al.
2/6



Note: picture image [205], used to illustrate the invention, is part of a vector image from CorelDRAW®9 which is protected by the copyright laws of U.S., Canada and elsewhere. Used under license.

Figure 2

FR 9 2000 0021
INCERTIS CARRO et al.
3/6

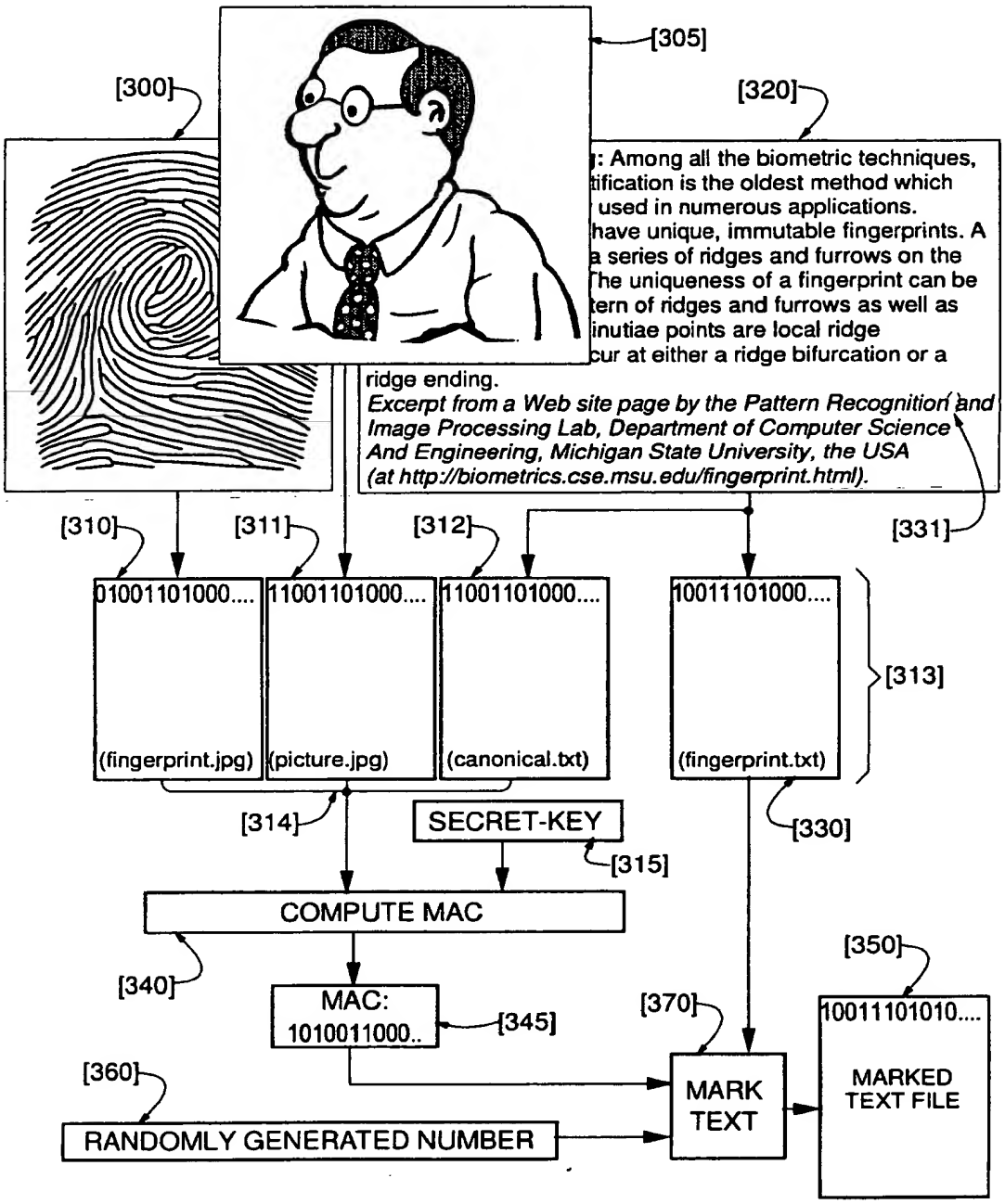


Figure 3

FR 9 2000 0021
INCERTIS CARRO et al.
4/6

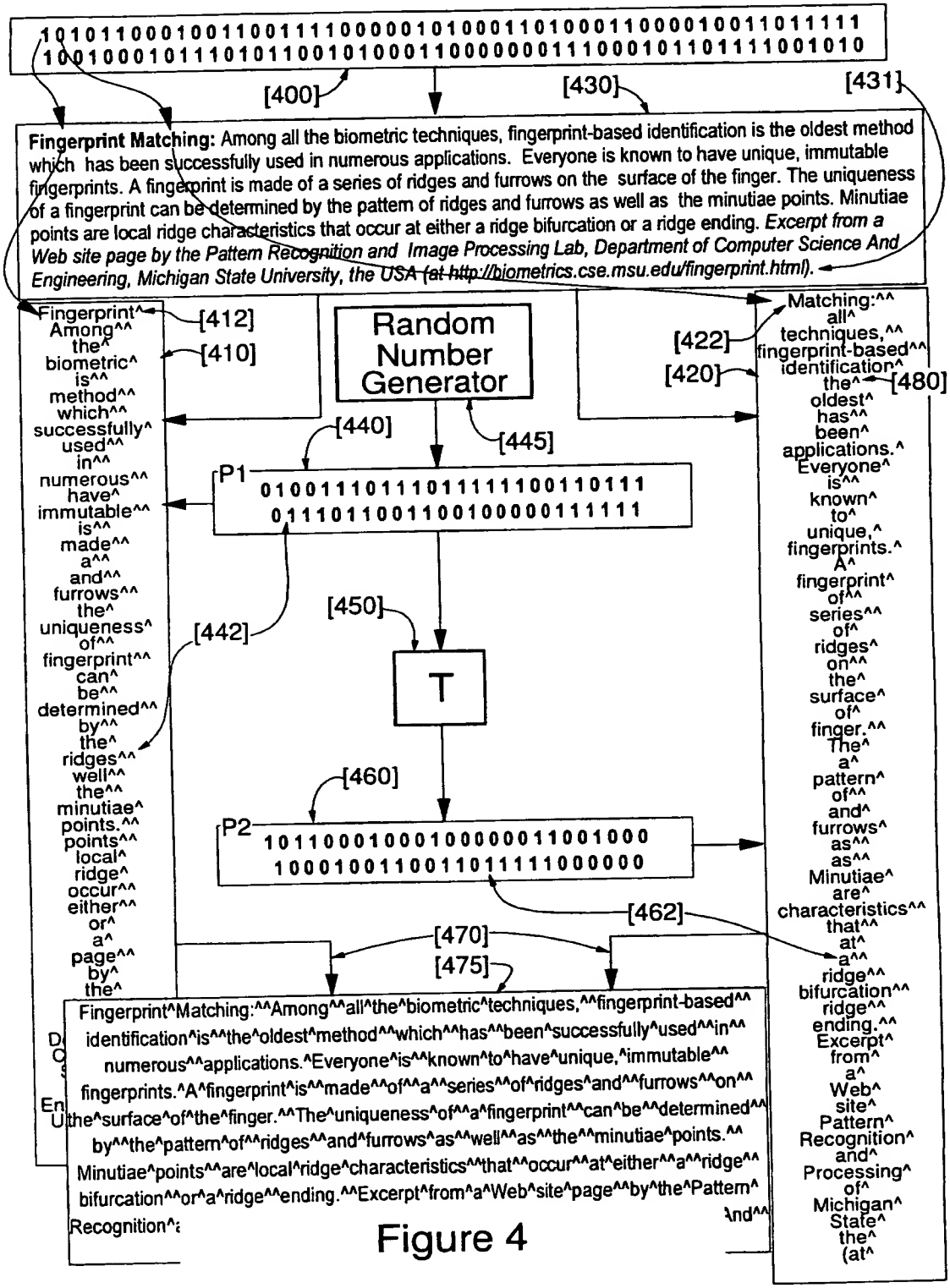


Figure 4

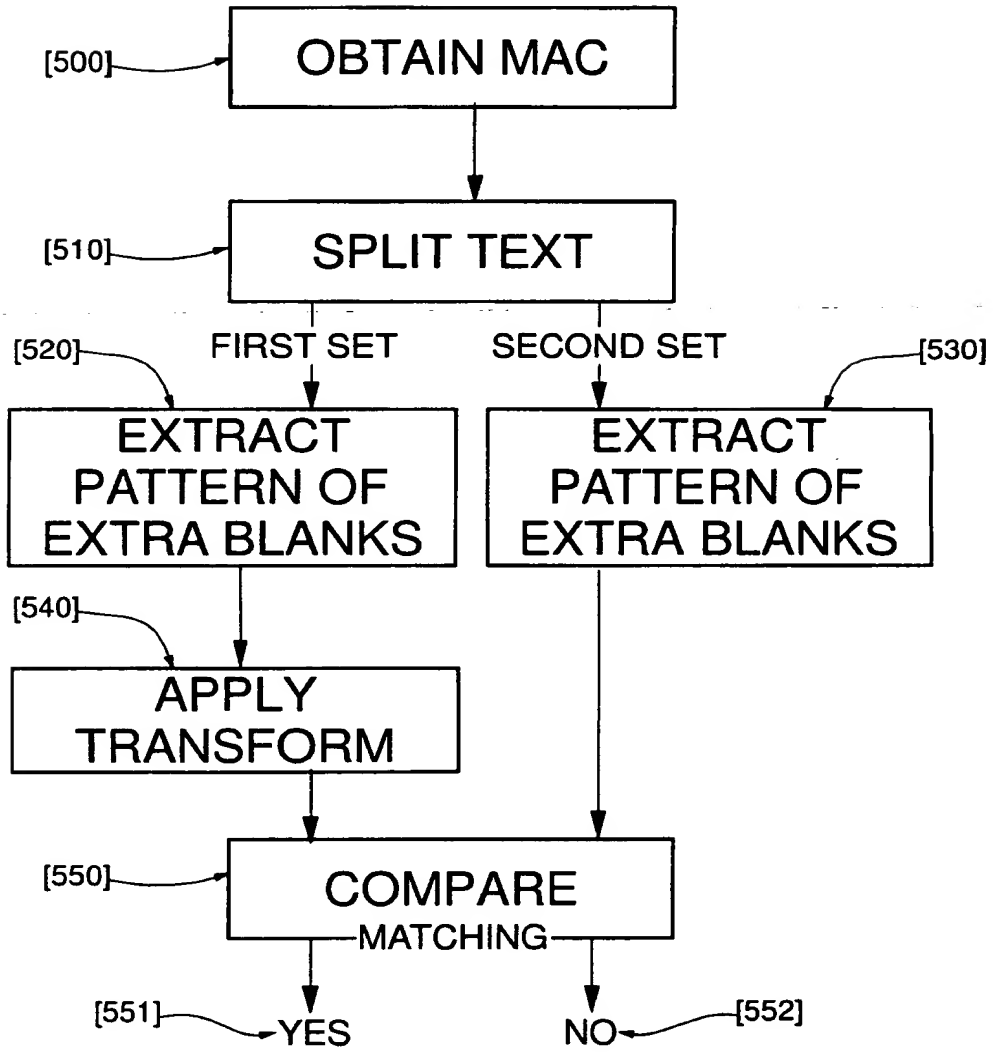

[500]
OBTAIN MAC
[510]
SPLIT TEXT
FIRST SET
SECOND SET
[520]
[530]
EXTRACT PATTERN OF EXTRA BLANKS
EXTRACT PATTERN OF EXTRA BLANKS
[540]
APPLY TRANSFORM
[550]
COMPARE
MATCHING
[551]
YES
NO
[552]

Figure 5

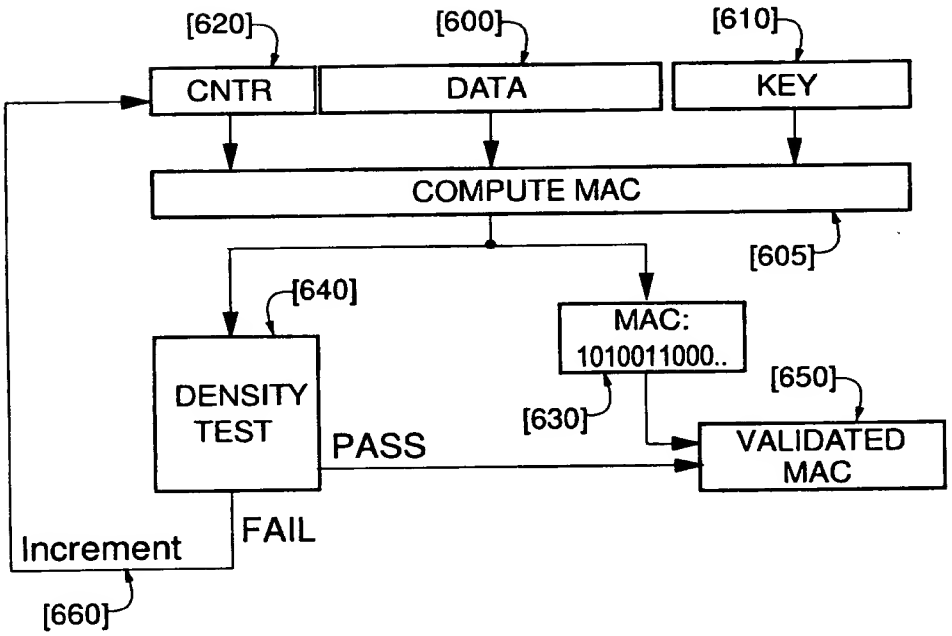

[620]
[600]
[610]
CNTR
DATA
KEY
COMPUTE MAC
[605]
[640]
MAC:
1010011000..
[650]
DENSITY
TEST
[630]
PASS
VALIDATED
MAC
FAIL
Increment
[660]

Figure 6